www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 447-453

e-ISSN : 2583-1062
Impact Factor : 7.001

ANALYSING EXTRACTION PROCESS INFLUENCE IN BIG DATA MINING AND A PROPOSED HYBRID MODEL

Priyansh Katiyar[1], Anshika Sahu[2], Dr. Shalini Lamba[3]

[1,2]Scholar, National Post Graduate College, Lucknow, India.

[3]Assistant Professor, Department Of Computer Science, National Post Graduate College, Lucknow, India.

DOI: https://www.doi.org/10.58257/IJPREMS44106

ABSTRACT

During the recent time of big data, a huge volume of unstructured and heterogeneous data is generated from various sources such as audio, video, text and images. In other words, the process of retrieving useful information from big and complex datasets is known as Big Data Mining. In the development of a big data system, the extraction process is an essential aspect as it affects the accuracy, reliability, and efficiency of decision-making. This paper involves the study of multiple data extraction methods and examines how they affect the big data mining. The major part of our research is that we have combine rule-based methods and machine learning (ML) and have created a hybrid approach which will enhance the accuracy and scalability while handling noisy and unstructured data. In addition, discussion the challenges of big data extraction, such as data quality issues, integration across diverse sources, and computational limitations. Optimization strategies are also reviewed. This research highlights an optimized and hybrid extraction strategy which increase data reliability and supports more accurate decision-making in big data infrastructure.

1. INTRODUCTION

Big data mining means a process of extracting some valuable information and insights out of huge and complicated collections of data. This may include the deployment of sophisticated analytical methods and instruments to find a pattern and trends amongst the data. The implications of these insights are wide spread in many fields like business intelligence, predictive modelling, and fraud detection. It is a subset of data mining, but it focuses specifically on large and complex data sets that may be too big to be analysed using traditional data mining techniques.

Big data mining is getting increasingly important in modern society as the large volume of data being generated continues to grow at an unprecedented rate. Then are some crucial ways in which big data mining is utilized-

1.1 Fraud discovery: Big data mining helps identify unusual sale patterns that could indicate fake action. This will help organisations to spot and help from fraud more snappily and effectively.

1.2 Healthcare: Big data mining analyses patient data, medical records, and inheritable information to prognosticate complaint outbreaks, epitomize treatment plans, and ameliorate patient issues.

1.3 Marketing: Companies dissect social media relations, reviews, and comments to gauge public sentiment, track brand character, and upgrade marketing strategies.

1.4 Education: Educational institutions dissect data on student’s performance, attendance, and engagement to identify at- threat scholars.

2. TYPES OF DATA SOURCES

2.1 Structured Data

Structured data is the type of data which is well defined, organized and stored in a predefined format or layout which make it easy to search, recover or examine. It is mainly stored in relational databases (in the form of tables) and spreadsheets.

Extracting structured data can cause complexity in extraction. While handling large datasets can leads to query performance issues. Moreover, ensuring data integrity and consistency across multiple sources can lead to faults in analysis and decision-making.

2.2 Semi-Structured Data

Semi-structured data is another type of data which does not follow any strict representation but still contains some level of organizational elements like tags and metadata. Common examples include emails, log files and NoSQL databases.

Semi-structured data is more flexible than previous type but needed processing before any analysis.

Inconsistency in data formats makes integrating complex and difficult. It requires parsing and transformation. Moreover, real-time sensor and log data faces scalability problems which demands efficient handling.

2.3 Unstructured Data

Unstructured data is a data which do not have any fixed format or structure that make it most difficult to process or analysis. It includes large text, images, videos, social media posts and IoT data. Unstructured Data denotes the majority of physical-world data.

Challenges in extracting unstructured data include required high processing power for processing images, videos, and audios. Additionally, ethical and privacy concerns with social media and medical data extraction.

3. EXTRACTION TECHNIQUES

Effective extraction techniques or methods are critical for managing the volume, variety and velocity of big data, ensuring that meaningful insights can be obtained with minimum redundancy. Different types of data mining techniques or methods which are used to guess desired output.

3.1 Classification

Classification is a widely used method for data extraction in big data mining, that helps in extracting meaningful data or information by arranging data into predefined groups. In simple words, classification is a technique of categorizing data points into predefined categories. Practical uses include text classification, Image recognition, facial recognition and anomaly detection.

3.2 Clustering

Clustering is an unsupervised technique used in big data mining to group same kind of data points based on patterns. As an extraction method, clustering helps identify, retrieve or extract meaningful data from large datasets without need of any predefined categories or labels.

Clustering is grouping similar data together without knowing what the groups are in advance. Example, automatically grouping news documents into different topics whereas classification is used to categorized data into pre-defined groups based on past examples. Example, classifying emails as spam or non-spam by previous categorized emails.

3.3 Regression

Regression is method mainly used for predicting numerical value as obtained data is used to predicting a continuous quantity for new observations. Unlike, classification means categorizing data or clustering means grouping data, regression are used to identifies relationships between variables and extracts expressive patterns from large amount of data.

Regression is used in big data mining for the reason that it aids in analysing relationship, predict future outputs and extract meaningful understandings from big dataset. This lets organization or businesses to extract key information and make better decisions.

Challenges of Regression method in data extraction are this method may not work well if relationships between data points are complicated or nonlinear, unexpected variations in data can affect accuracy and noisy or incomplete data can reduce extraction efficiency.

3.4 Outlier Detection

The word ‘outlier’ is a database that contain data objects which are significantly different from the rest of the dataset. These data objects do not follow the normal pattern and stands out because it is widely different from other values. The study of outlier data is known as outlier mining. Outlier Detection is the method which is used to extract and identify unusual, uncommon, anomalous and rare data points that differ from rest of the data. Outliers matter because it detects errors, identify fraud, discover trends and detect anomalies.

But many outliers can be valuable insights while some may be errors. Advanced computational techniques are required for processing big data for detecting outlier. These are some challenges of this method.

3.5 Sequential pattern

Sequential pattern extraction refers to the process of discovering regular and meaningful sequences of procedures, transactions, or actions in big datasets. This method is different from traditional data mining, as traditional mining concentrate on individual data points whereas sequential pattern mining focuses on the order in which events occur. Issue related to scalability, Noise & Missing Data or High Dimensionality i.e. large feature sets create pattern extraction complex and Pattern Overload.

4. HOW DATA EXTACTION PROCESSES INFLUENCE THE BIG DATA MINING

The accuracy, efficiency and reliability of data insights are directly influenced by the extraction process which plays a vital role in big data mining. An effective extraction method ensures the high-quality, scalable error free data, while a non-structured or ineffective method can lead to errors, in consistencies, and inefficiencies which reduced the reliability in big data analytics. The following are areas where extraction processes influence big data mining:

4.1 Quality and Accuracy of Data

The efficiency of any data system depends on quality of fetched data. Appropriate extraction processes eliminate the inconsistency and errors ensuring that the extracted data is reliable and accurate. Poor extraction process may cause faulty analysis, unfair insights and wrong decision-making.

4.2 Scalability and Performance

Processing of large datasets that grows continuously is known as big data mining. Big data mining requires processing terabytes or petabytes of data. Effective extraction techniques optimize data retrieval speeds, decreases the latency and storage overhead. On other side, ineffective processes can cause slow down real-time analysis and stress computational resources.

4.3 Effect on Machine Learning and AI

AI and machine learning models needed well-structured and high-quality datasets to produce accurate predictions. A powerful extraction model ensure that data cleaned and labelled correctly as well as formatted which enhances model performance. Weak extraction methods can raise noise, inappropriate features, or lost values, leading to biased models and wrong results.

4.4 Data Integration Across Sources

Big data is mostly collected from numerous structured and instructed sources, such as databases, social media, cloud platforms and IoT devices. A strong extraction process aids seamless data integration and allowing for comprehensive analytics. If extraction fails to extract correct data from multiple formats, may lead to inconsistencies and redundancies, in analytical accuracy.

5. CHALLENGES IN THE DATA EXTRACTION PROCESS

5.1 Issues Related to Data Quality

In data extraction process extracting high data quality is a big challenge, this challenge arises from old- fashioned or primitive information, data entry mistakes, or inconsistencies across different sources, to make the information collected is accurate and dependable the extraction process must include severe data satisfaction and confirmation ways. If these criteria not satisfied, integrity of posterior analyses can be compromised which may lead to defective perceptivity and opinions. So, it is important to extract high quality data.

5.2 Volume Of High Data

There are various challenges faced during the extraction of large volume of data. In today's modern world with increase in demand and supply chain it has become difficult to maintain data, as organizations now generate and collect more data, and the process of extraction becomes slower and more complex, this leads to increased processing times, system bottlenecks, and even data outages, where the system is unable to handle the demand. Caching and load balancing can mitigate the some of these issues.

5.3 Compliance and Security Considerations

Legal Data extraction landscape became complicated by legal and moral thoughts or considerations. While handling sensitive information, compliance with data privacy regulations is very important. Organizations must implement strategies like data anonymization and consent management to ensure they meet legal requirements while extracting data. Failure to obey to these regulations can cause legal penalties and reputational damage.

5.4 Technical Limitations

Lastly, extraction tools and techniques are constrained technically, which is a big challenge. Several available tools might not enable the combination of more sophisticated methods, including API extraction or machine learning algorithms, which are being demanded more and more to enable efficient data extraction in complex conditions. The absence of the right tools may put constraints on the capability of an organization to tap and leverage data effectively.

6. OPTIMIZATION TECHNIQUES FOR EXTRACTION

Given the accelerating size and complexity of big data, effectively performing meaningful information mining is a major challenge. Performance, accuracy and scalability weakness of old methods causes ineffective data managing. In order to solve these limitations, some advanced methods which can achieve efficient data extracting are used.

To overcome these challenges, several advance techniques are used that improve data extraction performance. Optimizing extraction process increases accuracy, reduces latency, and helps in better decision-making. There are following techniques:

6.1 Automation of Extraction Process

Automation the extraction process means automatically identify, retrieve and transform data from different sources into a structured format by using technology and software which minimize human interactions. Non-automatic extraction is ineffective, have errors and time-taking for large datasets. Automating extraction processes using tools such as ETL (Extract, Transform, Load), Web Scrapers and APIs (Application Programming Interfaces) decrease human interactions and increases efficiency. Automation also improves data integrity by reducing processing time.

6.2 Use of AI and Machine Learning in Optimizing Data Extraction

This is a critical area as it applies new technologies in improving the performance of computing and also in effective data processing in large and complex data. Artificial Intelligence (AI) and Machine Learning (ML) dramatically enhance data extraction through task automation and increased accuracy. These technologies enable systems that continue to learn from data pattern and improve in accuracy over time." Natural Language Processing (NLP) is used to extract useful summary information from unstructured text sources like emails, social media messages and news items. Computer Vision automatically extracts data from images, video, and scanned documents.

6.3 Distributed and Parallel Processing for Large-Scale Data Extraction:

The larger the quantity of data, the more sluggish and inefficient are the classic systems. This is addressed by distributed and parallel processing methods that accomplish the same by dividing large datasets into smaller and manageable pieces and processing them at the same time. The present-day big data extraction needs distributed and parallel computing that would make it possible to process it much faster, more efficiently, and scalable. Whereas, distributed systems are better at processing large volumes of information by using more than one machine, parallel computing specializes in operating only one system. They used together represent the backbone of high-performance big data analytics, powering innovations in AI, finance, healthcare, and real-time processing sectors. Together with them are the foundations of high-performance big data analytics, driving AI, finance, healthcare, and real-time processing innovations.

7. HYBRID APPROACH

With the growing significance of big data, the extracted data exists in various forms in structured (such as tables), semi-structured (such as XML/JSON), and unstructured (such as emails, PDFs, social media) forms. A single method (either rule-based or ML-based) often fails to handle all types efficiently. Rule-based method has disadvantage that they fail on noisy and unstructured data and needs manual updating. Conversely, ML models are relay on large labelled datasets and may misclassify rare patterns. A hybrid model integrates both techniques (rules-based and ML) is proposed as the solution for these problems.

Hybrid model uses rule-based extraction to handle high-confidence (degree of certainty or reliability) data while machine learning takes over when the data is ambiguous or the confidence is low. This double-layered architecture improves the extraction accuracy, adjusts effectively to type of data and reduces the chance of missed information. It offers a smarter and more scalable solution big data mining.

Steps involve in Hybrid Approach:

1. **Data Gathering:** Data is gathered and obtained in different sources.
2. **Preprocessing:** Remove the noisy data, deduplication, value normalization, and text tokenization for NLP models.
3. **Rule-Based Extraction:** Manual rules and patterns are used to extract information from structured or semi-structured data.
4. **ML-Based Extraction:** ML models are used to extract the relevant pattern when rule fails, or on unstructured data.
5. **Decision Integration:** relies on confidence levels to combine outputs, preference to high-confidence rule-based results but ML predictions when confidence in rules is low.
6. **Post-Processing:** The data that has been extracted is reconcilable, validated, de-duplicated and stored in an organized manner that is used in further analysis or mining.

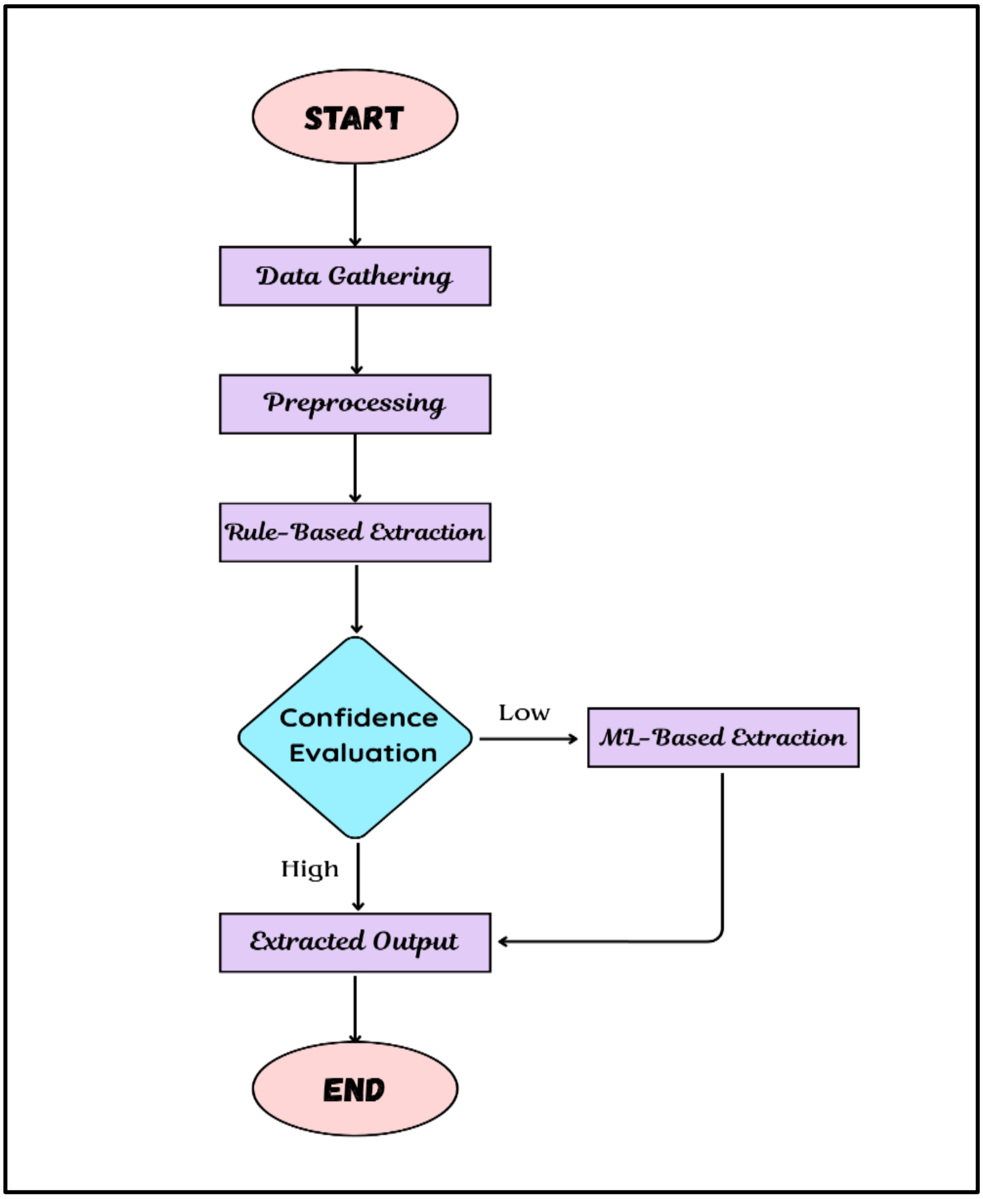
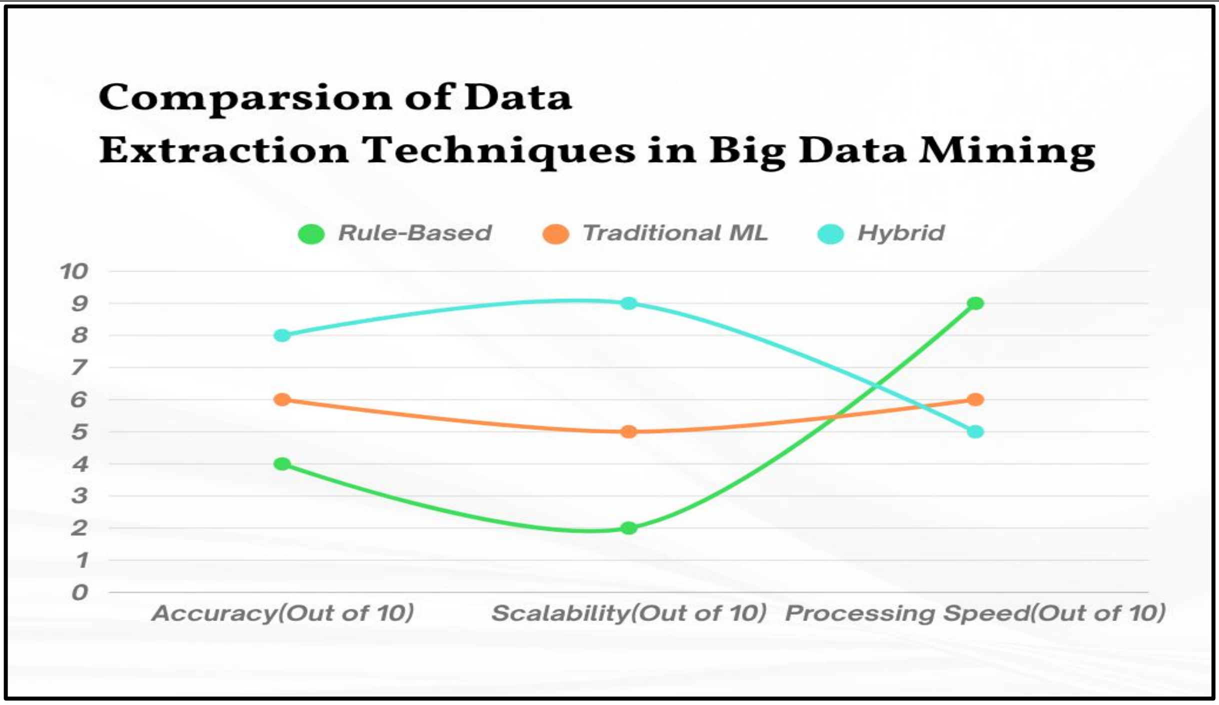


Figure 1: Flowchart for Hybrid Model

Table 1: This table is a side-by-side comparison of different extraction techniques.

Extraction Method	Approach	Accuracy	Scalability	Speed	Degree of Complexity
Rule-Based Extraction	It extracts based on predefined rules and patterns	Less	Poor scalability	Fast for small data	Simple
Traditional ML	Relies on trained models for classification and pattern detection	Moderate	Scalable for medium data	Moderate	Medium
Hybrid Methods (ML + Rule-Based)	Mix of rule -based patterns and ML techniques	High	Very High Scalable	Slow but works for big data	Medium



Graph 1: This graph is used to compare various extraction methods based on their accuracy, scalability and speed of processing.

Table 2:

Extraction Method	Accuracy (Out of 10)	Scalability (Out of 10)	Processing Speed (Out of 10)
Rule-Based	4	2	9
Traditional ML	6	5	6
Hybrid Methods	8	9	5

8. FUTURE WORK

Future innovation could involve the combination of deep learning models, blockchain technologies and large language models (LLMs) to enhance its performance, especially in unstructured text and image data areas. Besides, the feasibility of the suggested hybrid model implemented in practice with help of distributed systems, Apache Spark or Hadoop, and Python-based ML systems can be examined to confirm its scalability and real-time efficiency in the real-life application.

9. REFERENCES

[1] Sowmya, R., & Suneetha, K. R. (2017, January). Data mining with big data. In 2017 11th International Conference on Intelligent Systems and Control (ISCO) (pp. 246-250). IEEE.

[2] Asaad, R. R., & Abdulhakim, R. M. (2021). The concept of data mining and knowledge extraction techniques. Qubahan Academic Journal, 1(2), 17-20.

[3] Hariharakrishnan, J., Mohanavalli, S., & Kumar, K. S. (2017, January). Survey of pre-processing techniques for mining big data. In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-5). IEEE.

[4] Kumar, S., & Singh, M. (2019). A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem. Big data mining and analytics, 2(4), 240-247..

[5] Hassan, M. M., Gumaei, A., Alsanad, A., Alrubaian, M., & Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. Information Sciences, 513, 386-396.

[6] Müller, T. (2024). HYBRID AI FRAMEWORKS FOR BIG DATA PROCESSING AND PATTERN RECOGNITION. International Journal of Artificial Intellegence, Data Science and Engineering, 1(02), 33-39.

[7] Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. IEEE Access, 11, 10535-10562.

[8] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of business research, 70, 263-286.

[9] Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big data: Promises and problems. Computer, 48(03), 20-23.

[10] Verma, A., Kaur, I., & Arora, N. (2016). Comparative analysis of information extraction techniques for data mining. Indian Journal of Science and Technology, 9(11), 1-18.

[11] Nde, D. B., & Foncha, A. C. (2020). Optimization methods for the extraction of vegetable oils: A review. Processes, 8(2), 209.

[12] Jonnalagadda, S. R., Goyal, P., & Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. Systematic reviews, 4(1), 78.

[13] Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. Big Data and Cognitive Computing, 4(1), 1.

[14] Haoxiang, W., & Smys, S. (2021). Big data analysis and perturbation using data mining algorithm. Journal of Soft Computing Paradigm (JSCP), 3(01), 19-28.

[15] Ingole, P., Bhoir, S., & Vidhate, A. V. (2018, March). Hybrid model for text classification. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 7-15). IEEE.

[16] El Azzouzi, M., Bellafqira, R., Coatrieux, G., Cuggia, M., & Bouzille, G. (2024). Secure extraction of personal information from ehr by federated machine learning. Studies in Health Technology and Informatics, 316, 611-615.

[17] Hu, F., Dong, S., Chang, T., Zhou, J., Li, H., Wang, J., ... & Wang, X. (2023, October). FedEAE: Federated Learning Based Privacy-Preserving Event Argument Extraction. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 326-337). Cham: Springer Nature Switzerland.

[18] Nguyen, D. P., Yu, S., Muñoz, J. P., & Jannesari, A. (2023, November). Enhancing heterogeneous federated learning with knowledge extraction and multi-model fusion. In Proceedings of the SC'23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis (pp. 36-43).